Bayesian Nonparametrics for Motif Estimation of Transcription Factor Binding Sites

Philipp Benner
Max Planck Institute
for Mathematics in the Sciences
Inselstrasse 22
04103 Leipzig, Germany
philipp.benner@mis.mpg.de

Pierre-Yves Bourguignon
Max Planck Institute
for Mathematics in the Sciences
Inselstrasse 22
04103 Leipzig, Germany
pierre-yves.bourguignon@mis.mpg.de

Stephan Poppe
Max Planck Institute
for Mathematics in the Sciences
Inselstrasse 22
04103 Leipzig, Germany
stephan.poppe@mis.mpg.de

Abstract

Gene expression is controlled by a family of proteins called *transcription factors* (*TF*) that bind specific regions of the DNA molecule and thereby promote or block the transcription of nearby genes. We are interested in the identification of *transcription factor binding sites* (*TFBS*) from experimental (*ChIP-Seq*) data of primate cells. This is a challenging problem because TFBS that are bound by a certain TF may vary and there is little or no knowledge about the generating distribution (*motif*). Experiments provide us with a set of sequences that are enriched with instances of a motif but also contain several other patterns such as simple repeats or retroviral insertions. We use the *Dirichlet* (cf. Ferguson [1973]) and *Pitman-Yor processes* (cf. Pitman and Yor [1997]) to express our prior belief about the distribution of the number of clusters and their sizes. In addition, a structure analysis of the transciption factor provides us with a first prediction of the motif. To incorporate this knowledge, we use a base measure which is a mixture of a uniform prior and a prior based on the predicted motif. Results from ChIP-Seq experiments usually contain between $2,000,000$ and $5,000,000$ nucleotides, which makes the common Gibbs sampling approach very demanding. The approximation of the posterior is further complicated by the Pitman-Yor process prior, which requires many samples until the stationary distribution is reached. We also explore other process priors that are motivated from species sampling problems and obtain similar results using substantially less sampling steps. Moreover, we are interested in exploring alternative methods such as *variational Bayes* approaches. To incorporate information from homologous sites we present an efficient algorithm for the computation of Felsenstein's evolutionary model. It allows us to obtain analytical expressions of many posterior quantities.

Whether or not a TFBS is present in the proximity of a gene is important information for the understanding of gene regulatory networks. To identify TFBS in the primate genome for a given *protein of interest* (*POI*) we infer a model (motif) for potential binding sites from ChIP-Seq data. These experiments provide us with a set of sequences of approximately 500 to 1000 nucleotides that are enriched with TFBS. It is assumed that whatever repeated pattern is found might correspond to a

potential motif. A first difficulty is that each sequence may contain no or any number of binding sites. Besides, it is likely that a single motif might not fully characterize all binding sites (cf. Barash et al. [2003]). It is also well known that noncoding DNA regions in primates contain many repeated patterns and low complexity sequences (simple repeats, tandem repeats, interspersed repeats). The approximate length of the binding sites is known from the protein structure. We also have a tentative prediction of the motif that was derived from the structure of the protein, whose reliability is however afflicted with high uncertainty. Unfortunately, little is also known about the binding process itself and the diversity of sites that are bound by the POI. A high diversity is often expected for long transcription factors that bind up to 50 nucleotides and should be expressed by a low information content of the motif.

This motivates the application of the Dirichlet and Pitman-Yor processes. First, they allow us to search for an unknown number of repeated patterns. Second, we can use a mixture model as base measure which consists of a uniform prior and predictions of the motif from the protein structure. This allows us to ignore the predicted motifs if they poorly match the data and in addition to identify unrelated repeated patterns. Finally, we know that the data set is enriched with TFBS and the motifs should be among the largest clusters with a potentially low information content. We therefore want to place a prior on the number of sites that belong to a motif, given by the cluster sizes.

For the Dirichlet process, the number of clusters grows logarithmically in the number of observations and its *rich-gets-richer* property causes some of them to be large. On the other hand, the Pitman-Yor process produces clusters whose sizes follow a power-law distribution, which is why it is preferred for statistical models of language (cf. Goldwater et al. [2006]). For our application there is little knowledge about what process might be best suited for the distribution of repeated patterns, but experimental validations of our predictions will hopefully provide a deeper insight.

We have developed a continuum of priors (cf. Poppe and Bourguignon) that are motivated by species sampling problems (cf. Bunge and Fritzpatrick [1993], Zabell [2005]). The continuum contains priors that satisfy exchangeability including the Dirichlet and Pitman-Yor processes. We test one such prior on the ChIP-Seq data set and compare the results with those from the Dirichlet and Pitman-Yor processes.

To compute posterior estimates we use the collapsed Gibbs sampling scheme (cf. MacEachern [1994], Neal [2000]), where the Markov chain is defined on the latent partitions of the data set. The nonidentifiability of clusters is a well known problem for the sampling of mixture models. To obtain an unbiased posterior estimate we record in every sampling step all subsequences that have been assigned to the same cluster. The posterior then resembles a similarity measure for all subsequences of a given length. This approach is computationally very demanding for large-scale applications, such as ours. The sampling process is further complicated by the Pitman-Yor prior and the rich-get-richer property. Large clusters form already at the very first sampling steps and it takes many more samples until all clusters are filled with appropriate subsequences.

We use information from homologous sequences in related species to improve the identification of TFBS. To model the evolution of such sequences, the model of Felsenstein [1981] is used. We have developed efficient algorithms to compute the likelihood with respect to a given phylogenetic tree with Felsenstein's evolutionary model. The methods allow us to obtain analytical expressions of posterior quantities.

Our preliminary results show that with our method, which is based on Bayesian nonparametrics, we were able to identify repeated patterns within ChIP-Seq data. The Gibbs sampling approach however is too expensive to sample the whole data set. By excluding known repeated patterns we were able to reduce the data set by half, which allows us to obtain several thousand samples from the posterior distribution. Nevertheless, it is highly questionable whether excluding known repeated patterns is a valid approach. Accordingly, we require an alternative method, such as variational Bayes, to process the full data set.

References

Y. Barash, G. Elidan, N. Friedman, and T. Kaplan. Modeling dependencies in protein-dna binding sites. In M. Vingron, S. Istrail, P. Pevzner, and M. Waterman, editors, *Proceeding of the Seventh Annual International Conference on Computational Molecular Biology*, pages 28–37. ACM

Press, New York, NY, 2003.

J. Bunge and M. Fritzpatrick. Estimating the number of species. a review. *Journal of the American Statistical Association*, 88(421):364–373, 1993.

J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376, 1981.

Thomas S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

S. Goldwater, T. L. Griffiths, and M. Johnson. Interpolating between types and tokens by estimating power-law generators. *Advances in Neural Information Processing Systems*, 18:364–373, 2006.

S. MacEachern. Estimating normal means with a conjugate style dirichlet process prior. *Communications in Statistics B*, 23(3):727–741, 1994.

Radford M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

J. Pitman and M. Yor. The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25(2):855–900, 1997.

Stephan Poppe and Pierre-Yves Bourguignon. An inductive characterization of a generalized class of dirichlet measures (in preparation).

Sandy L. Zabell. *Symmetry and its discontents: Essays on the history of inductive probability*. Cambridge studies in probability, induction, and decision theory. Cambridge Univ. Press, 2005.